

COVER PAGE

Document downloaded by @DAEL

Wed Nov 11 23:48:37 2026
For personal use

When automatic English translation is provided, only the original document is authentic.
The EAA cannot be held responsible of any translation error

Bibliographical reference

Redrawing the Link Between Customer Satisfaction and Speech Quality, Laetitia Gros, Virginie Durin and Noël Chateau, Acta Acustica **vol. 94** (Number 1), 2008, pp. 32-42

DOI
https://doi.org/10.3813/AAA.918005

ACTA ACUSTICA UNITED WITH ACUSTICA
Vol. 94 (2008) 32 – 42

DOI 10.3813/AAA.918005

Redrawing the Link Between Customer Satisfaction and Speech Quality

Laetitia Gros, Virginie Durin, Noël Chateau
France Telecom, R&D Division, Technologies R&D Center. laetitia.gros@orange-ftgroup.com

Summary
This paper analyzes the actual use of Mean Opinion Scores (MOS) in the telecommunication industry and looks back on the tremendous evolution that has encountered this sector since subjective test methodologies were standardized by ITU-T (International Telecommunication Union-Standardization). It is pointed out that MOS have severe drawbacks and that test methods should evolve by taking into account valuable inputs from connected research fields such as the psychology of emotions, cognitive science, usability engineering and marketing research.

PACS no. 43.66.Lj

1. Introduction

In a highly competitive market, price and quality of services are two key issues that telecommunications operators have to take care of. Whereas price issues are generally only handled by marketing units, those on quality are the concern of many units, from marketing, networks, to Research & Development ones. Research & Development centers, which have a long experience in running subjective tests aiming at assessing speech or video quality of technological components (such as codecs, noise-suppressor or other signal-processing algorithms) are now asked by marketing units to produce data that reflect customers' opinions on quality and their subsequent satisfaction in their daily use of services. As a consequence, the 'famous' MOS (*Mean Opinion Score*) is not restricted anymore to research labs, but is extensively used by marketers.

Figure 1 from ETSI (European Telecommunications Standards Institute) [1] shows the trends of end-to-end quality in speech telecommunications as a function of technological progress. Till the end of the eighties, the different technological improvements in the field of telecommunications were synonymous of improved quality delivered to the users. Mobile networks and packet voice have definitely introduced two revolutions, one on uses, the other one on pricing and added services, but they have also introduced new uncontrolled speech impairments.

Before the introduction of these technologies and the apparition of competition on the European market that both occurred in the nineties, speech quality of voice communications was guaranteed and rather homogenous among users and over time. Moreover, the use was restricted to land-line communications, only few terminals were available, and the pricing was fairly simple and the same for all users. Thanks to this homogeneity of quality, use, terminals and pricing, the MOS measured in research labs were considered as good predictors of users' opinion and satisfaction of the service, since almost no other criteria would come into the scene. Today, we have to face a strong heterogeneity of quality, uses, terminals and pricing. In such a context, it is probable that the MOS measured in labs do not reflect as well as thirty years ago users' (who became customers) opinion on the quality of the voice communication services, and therefore their satisfaction.

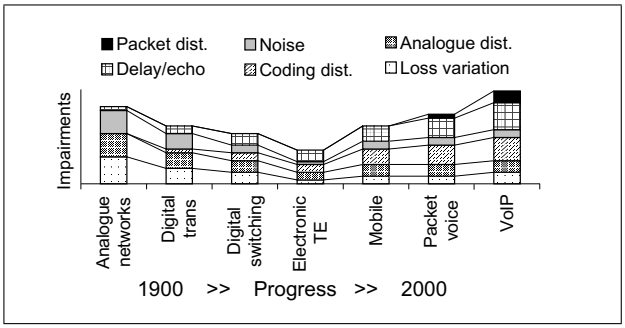


Figure 1. Speech quality impairments as a function of technological progress.

This paper reviews some of the major drawbacks of the extensive use of MOS in today's context and proposes to consider speech quality in a broader context, including emotional reactions and behavioural adaptations of users of voice telecommunication services facing low speech quality periods. First results on the impact of speech degradations on behaviour and on cognitive processes through performances are given. Then inputs from the research field of the psychology of emotions, and notably from the appraisal theory developed by Scherer [2] are discussed in order to propose a refined definition of customer satis-

Received 1 March 2007,
accepted 12 November 2007.

© S. Hirzel Verlag · EAA

Document downloaded by @DAEL, 2026-11-11 23:48:40 - personal use only

faction in the field of voice telecommunication services. Methodological implications are discussed and directions for future work are proposed in the last section.

2. Three major drawbacks of the Mean Opinion Score

The MOS is undeniably a very useful tool for diagnosing and assessing technological components of voice communication services [3]. MOS generally refer to Mean Opinion Scores obtained by averaging individual quality scores given by listeners on the five-category scale (5 = Excellent, 4 = Good, 3 = Fair, 2 = Poor, 1 = Bad) for short speech samples processed by the system under study (Absolute Category Rating-listening tests) [4]. However, MOS can also be established in conversation tests [4]. Additionally, the development of objective measurement tools such as PESQ (Perceptual Evaluation of Speech Quality) [5] has extended the concept of MOS to the field of objective quality. Although there is a wider and wider use of MOS, we wish to point out three of their major drawbacks.

2.1. MOS and the quality of voice communication services

As discussed in the introduction, MOS were rather convenient before the nineties for predicting users' opinion on the global quality of voice communication services. After the mobile and the VoIP revolutions, voice communication services have become complex services, including new functionalities (e.g. messaging, directory, chat), being accessible from a wide variety of terminals (fixed, cordless, mobile), through various speech qualities (associated to landline, mobile, VoIP networks, their specific codecs and all the tandeming possibilities), in various contexts of use (at home, in the streets, in public/private transportation, etc.). This explosion of heterogeneity has blurred the link between speech quality, as measured in laboratories, and service quality, as perceived by users/customers. However, there is a growing demand of marketing and network units for producing speech quality data, as perceived by customers, in order to "market quality" for the first, and to monitor it on networks for the second. Such an extensive use of MOS data and their related extrapolations are really dangerous, since many non-experts use them "as the truth", reflecting end user's opinion, which is of course not the case. Although one way of limiting the problem is obviously communicating with pedagogy on the limits of the interpretation and extrapolation of MOS data, there is also an urgent need for developing new methods allowing to producing data that are closer to end users' perception of service quality.

2.2. Corpus dependence

MOS values are obtained by averaging ratings on a category scale. Such scales have a rather serious defect since the data they produce are dependent on the corpus distribution. The bias comes from the fact that whatever the corpus is, subjects tend to assign stimuli to categories in such a way that all categories are used about equally often [6, p.108]. If the corpus contains an under or an over-representation of low or high-intensity stimuli, the psychophysical function is transformed by a non-linear function. In the case of speech quality assessment, the use of corpuses with only narrow band stimuli, or with mixed narrow and wideband stimuli typically exemplifies this problem. Barriac, Le Saout and Lockwood [7] have shown that narrow-band stimuli see a shift of their MOS between 0.01 and 0.63 when the narrow-band corpus is extended to a mixed narrow and wideband one. Consequently, the ITU-T recently adopted a new Mean Opinion Score terminology, in order to take into account the three narrow-band, wide-band and mixed-band contexts [8]. In the context of deployment of wideband speech in VoIP services (e.g. Orange HD and AOL in France, BT Broadband in UK, Skype and MSN worldwide), the instability of the MOS scale is rather critical to clearly compare quality levels.

2.3. Explicit assessment

Up to now, most of researchers have considered sound quality as an attribute of sound that can be expressed in terms of auditory sensations or by a hedonic judgment [9]. In many studies, sound quality is either characterized by the qualification/quantification of the perception of the acoustic features of the stimuli (e.g. [10]) or by hedonic judgments such as the pleasantness of the stimuli (e.g. [11]) or their disturbance/annoyance (e.g., [12]). All these sound and speech quality approaches are based on an explicit judgment, which is not likely to take place in real life.

In ecological situations, speech quality is generally not a conscious object for users, as they focus their attention on the (para-) linguistic content of the communication. However, when the frequency and magnitude of degradations reach an amount that affects the communication (e.g. need to concentrate, to repeat, to interrupt), speech quality becomes clearly conscious for users and even a topic of the discussion. To make speech quality conscious and to be assessed, it has to become "a topic for thought" through the process of reflection [3, p.55]. For Merleau-Ponty, this "intellectualization" of quality is a mistake [13, p.11] that he describes by the following sentence: "we consider quality as an object *of* consciousness whereas is it an object *for* consciousness."

As a consequence, it might be argued that if one wants to study the object "speech quality" as closer as possible to what it represents for an end user in an ecological communication situation, he/she might not directly ask the user's opinion on speech quality but find alternative methods. We propose that speech quality can be indirectly analyzed through its impact on user's behaviour.

3. Speech quality and users' behaviour

Speech quality is one of the elements of quality of the telecommunication system/service. If it is high, users will

Document downloaded by @DAEL, 2026-11-11 23:48:40 - personal use only

Table I. Example of a calculation set achieved with the twelve instructions and the start one.

Start instructions: "As start value, take 15 (or 20)"	
1.	"Add the number 3."
2.	"Then remove 8."
3.	"To this result, add the number 10."
4.	"Now, add 4."
5.	"To this result, please remove 9."
6.	"Now please add 5."
7.	"Now, to the computed number, remove 1."
8.	"To the temporary result, add 12."
9.	"From the temporary, subtract 7."
10.	"To the current result, add 11."
11.	"To the previous result, remove 2."
12.	"Please, subtract 6."

not encounter specific problems to orally communicate with their interlocutor. It is highly probable that they will not even pay attention to speech quality and that it will not be an explicit object of their consciousness. On the contrary, if it is low, below a certain threshold, it will become a conscious object and will impact their behaviour. Typical behavioural adaptations to low speech quality will be for example increasing concentration, asking the interlocutor to repeat or to speak louder, walking away (from a noise source or to find a better network reception), terminating the communication and trying a few seconds later, etc. These behaviours can be essentially cognitive (e.g., increasing concentration, that might only by expressed by a facial mimic such as frowning eye brows) or essentially motor (e.g. walking to find a better reception for mobile phones) or both (e.g., deciding to interrupt the communication and dialing again immediately after). These few examples clearly show that the observation of users' behaviours might be a valuable source of information on speech quality and on its impact on subsequent users' satisfaction of the service. Behind these evident and conscious behaviour adaptations to important speech quality deterioration, it may be assumed that more subtle, indeed not conscious behaviour "variations" occur when speech quality impairs.

Mullin, Smallwood, Watson and Wilson [14] explored how to measure the audio and the video quality with the use of electrophysiological signals. The considered criteria were the Galvanic Skin Response (GSR), the Heart Rate (HR) and the Blood Volume Response (BVP). Such signals are known to directly reflect subjects' stress and emotion [15]. Results showed that different electrophysiological responses could be obtained for different degradations, and could be partially dependent of the task. Therefore, electrophysiological signals might be useful to assess the impact of quality on users' stress and emotion.

Sonntag, Portele and Haas [16] measured the understanding of speech sequences uttered by six speech synthesis systems and with a natural voice, coded or not by the GSM network (so likely to be impaired) in a dual task. The primary and the secondary tasks both showed significant differences in reaction times for the different types of voices. In addition, for two of the seven voices, the differences between the coded and the non coded differences were significant. Therefore, it seems possible not to ask subjects about their conscious perception of speech quality, but to observe, measure and characterize how speech quality influences his/her behaviour in task-oriented protocols.

Two studies were conducted in order to understand how speech quality impacts performances and behaviours in different tasks involving auditory modalities with impaired speech signal. The first one [17] was based on a dual task and largely inspired by the study of Sonntag, Portele and Haas [16]. The second one was based on a simple task, based on the dichotic listening paradigm.

3.1. Assessment of the impact of speech quality on performance in a dual-task protocol

3.1.1. Methodology

The paradigm used in this experiment is a dual task: the primary task (a mental computing exercise) involved the auditory modality using speech signals with different levels of quality. A secondary task, involving the visual and motor modalities, was used to overload the cognitive charge, and to create a stress state for the subject.

Table I gives an example of a calculation set used in the primary task. Different calculation sets were prepared with different orders of presentation of these twelve instructions and two start instructions. Therefore the speech material was exactly the same for all calculation sets.

The fourteen instructions (twelve instructions and two start instructions) were uttered by a female speaker and recorded in a sound proof room at 48 kHz. They were synthesized with the France Telecom's CVOX Text-To-Speech system (female voice, diphone version) at 16 kHz. Then the twenty-eight resulting speech signals were downsampled at 8 kHz, and degraded with noise using the Modulated Noise Reference Unit (MNRU), recommended by the International Union [18], at a level Q of 15 dB (Q being the ratio of speech power to modulated noise power). In addition, the fourteen speech signals with natural voice were degraded by simulating losses of signal packets (10% of packet losses with a uniform law), as encountered in the internet network. The signal was coded with a G.723.1 codec (11-ms frames per packet, 3 frames per packet). Thus, there were five speech quality conditions: natural voice (N), synthetic voice (S), natural voice with noise (Nn), synthetic voice with noise (Sn) and natural voice with packet losses (Nip).

Twenty naive subjects participated to the test which proceeded in two sessions lasting approximately 15 minutes each. In each session, subjects computed five calculation sets (with five different instruction orders) with one quality condition per set (the quality condition order was random and different for all subjects). The start value was 15 in the first session and 20 in the second one.

Document downloaded by @DAEL, 2026-11-11 23:48:40 - personal use only

For each set of calculation, subjects were asked to mentally compute and to orally give the answer after each instruction, as quickly as possible. In the same time, they were asked to click with a mouse on the square, among four colored ones appearing on a screen, whose color matched a randomly displayed color on the monitor. At the end of each calculation set, subjects were asked to give their opinion on speech quality on the five-category MOS scale [4]: Excellent (= 5), Good (= 4), Fair (= 3), Poor (= 2), Bad (= 1).

Twelve reaction times RTs (time between the end of the instruction and the beginning of the subject's answer) were collected for each calculation set, so twenty-four RTs per quality condition (cf. two calculation sets per quality condition). RTs for the secondary task (time between the appearance of the colored square and the subject's mouse click) were also measured. For the two tasks, the number of errors done by subjects in their calculations is computed.

The experimental design was a factorial design with two fixed factors (factor Quality with five levels and factor Instruction with twelve levels) and one random factor (factor Subject with twenty levels).

In addition, during the whole test, the Galvanic Skin Response (GSR) was measured with two electrodes fixed at two fingers of subjects' free hand. Skin conductivity represents changes in the sympathetic system which role consists in triggering reactions such as a rush of adrenalin faced with a potential danger. Skin indicator is a good indicator of stress and arousal; under stress, the GSR increases [15].

3.1.2. Results

Figure 2 shows the mean opinion scores (MOS) and the associated 95-% confidence intervals obtained for each quality conditions, averaged on all the subjects. As excepted, the quality judgments as measured by the typical MOS confirm that the different experimental conditions were perceived by subjects as distinct speech quality levels, from poor to good quality (confirmed by an ANOVA: $F(4,76) = 28.3$, $p < 0.001$).

An ANOVA conducted on the RT data considering the Quality (Q), Instruction (I), and Subjects (S) experimental factors shows no significant effect of the Quality and Instruction factors; in return, a strong effect of the Subject factor was found ($F(19,240) = 29.01$, $p < 0.001$). In order to remove the variability due to the inter-subject differences, the individual RTs were centered and reduced. An ANOVA was conducted on the resulting RTs' z-scores. Again, no significant influence of the speech quality levels was observed ($F(4,960) = 1.17$, $p = 0.32$). In addition, the distribution of the frequency of errors was not significantly influenced by the speech quality of instructions.

Now Figure 3 shows RTs' z-scores and the associated 95-% confidence intervals for the five speech quality conditions, obtained in the secondary task. Contrary to the primary task's RTs, it seems that secondary task's RTs are dependent on speech quality in the sense that the more degraded the quality, the longer the reaction times. An ANOVA was conducted and showed a significant effect of the Quality factor ($F(4,76) = 3.84$, $p < 0.01$).

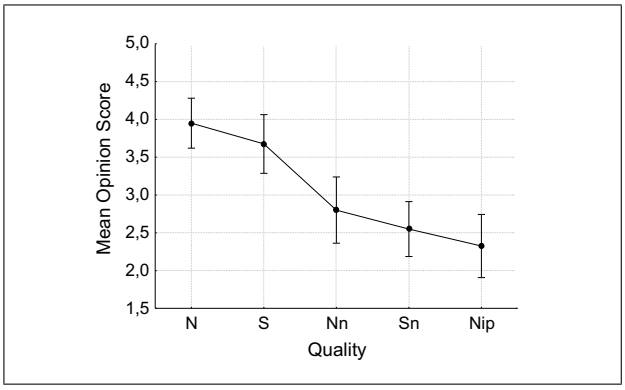


Figure 2. Mean Opinion Scores obtained for the different speech quality conditions.

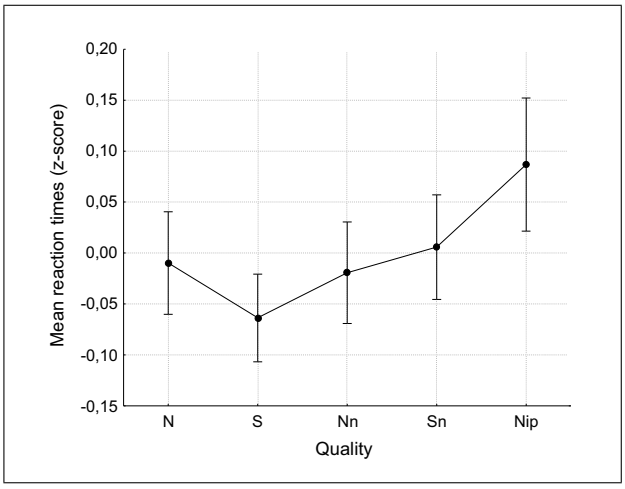


Figure 3. Reaction times' z-scores according to the speech quality conditions, for the secondary task.

Contrary to Sonntag, Portele and Haas [16], we did not find differences in RTs for the calculation task between natural and synthetic voices. However synthetic voices have improved during the last seven years (diphone-based speech synthesis being replaced by corpus-based speech synthesis since 1998), and the TTS system used for our experiment provided high-quality speech signals (MOS = 3.68), close to that of natural voice (MOS = 3.95).

The speech quality factor did not influence RTs of the primary task but significantly influenced RTs of the secondary task. Concerning the primary task, mean RTs are long (as compared for instance to RTs measured for the detection of audio signals which are around 160 ms [19]) as they result from complex and resource-demanding cognitive processes (mental calculation). Their variability is probably the consequence of the variability of the complexity of the various calculations subjects had to achieve (e.g., 21-12 being more difficult than 21+1). Therefore, the potential influence of speech quality of audio instructions on RTs may have been hidden by such large variations.

Document downloaded by @DAEL, 2026-11-11 23:48:40 - personal use only

Table II. Conditions combining degradations (MNRU at 25, 15 or 5 dB or Bandwidth BW of 300-3400 Hz, 50-7000 Hz or 50-14000 Hz) and listening situation (Asymmetrical or Symmetrical). RALB: Right Ear for group A, Left Ear for group B, LARB: Left Ear for group A, Right Ear for group B.

Trial	Situation	RALB	LARB
1	Asymmetrical	HQ	HQ
2	Asymmetrical	HQ HQ	MNRU 25 or BW 50-14000
3	Asymmetrical	HQ HQ	MNRU 15 or BW 50-7000
4	Asymmetrical	HQ HQ	MNRU 5 or BW 300-3400
5	Symmetrical	HQ	HQ
6	Symmetrical	MNRU 25 or BW 50-14000	MNRU 25 or BW 50-14000
7	Symmetrical	MNRU 15 or BW 50-7000	MNRU 15 or BW 50-7000
8	Symmetrical	MNRU 5 or BW 300-3400	MNRU 5 or BW 300-3400

However, for the secondary task, the worst speech quality (Nip, MOS = 2.3) highly lengthens the RTs. A detailed analysis of subjects' behaviour showed that, when mentally calculating (primary task), they tended to suspend their activity related to the secondary task till they orally gave their answer. This behaviour suggests that tasks were not treated in parallel but sequentially. It seems that for the worst speech quality, this suspension strategy is amplified. Such behaviour can be related to a bottleneck model of dual-task processing. When the main task requires more cognitive resources (in our experiment, because of a worse speech quality), the attribution of additional resources is done to the detriment of the secondary task [20]. Therefore it can be hypothesized that speech quality more impacts the performances and management strategies of non-auditory subsidiary tasks than those of auditory main communication tasks.

As for RT data, GSR individual data were transformed into z-scores in order to remove inter-individual variability. Due to technical problems, only the data of the first calculation set of each subject could be measured. An ANOVA was conducted on GSRs' z-scores. The factor Quality does not significantly influence subjects' GSR ($F(4,76) = 0.45$, $p = 0.77$), suggesting that the stress level generated by the accomplishment of the two tasks was independent on speech quality.

3.2. Assessment of impact of speech quality on performance in a dichotic-listening-task protocol

In the previous experiment, the use of a dual-task paradigm aimed to overload subjects when achieving a task involving impaired speech signals. However, results showed that this overload was limited by strategies specific to the dual task (for example, prioritization of the main task and processing of the secondary task when the main task required less resource). Therefore, a second experiment was conducted in order to explore the impact of speech quality on performance in a single task that required a high level of attention.

3.2.1. Methodology

In this second experiment, subjects were placed in a dichotic listening situation, that presented the advantage to share the attentional processes, and consequently to increase the cognitive load. Subjects listened to different lists of fifty dissyllabic words presented on the two ears (one list per ear), with different and irregular rates (then overlapping between presentations on left and right ears was possible). Among the one hundred presented words, twenty ones were pointed out as target (ten on the left ear, and ten on the right ear): they appeared successively on a PC screen face to the subject. Subjects had to detect these target words either on the right ear or on the left ear, and to indicate if the target word was on the right or on the left ear. Once detected, the following target word appeared on the screen.

Three levels of MNRU [18] (25, 15 and 5 dB) were applied to the words uttered by a male speaker. And three bandwidths BW (narrow band 300-3400 Hz, wide band 50-7000 Hz and super wide band 50–14000 Hz) were applied to the words uttered by a female speaker. Consequently, for each type of degradation (MNRU or bandwidth), four quality levels were considered, with the high quality level HQ (without any MNRU or filter) included. The lists of words were presented at 48 kHz and 73 dB SLP through headphones.

Twenty-four right-handed naive subjects participated to the test which proceeded in two sessions (lasting approximately 15 min each), one for the MNRU degradation, and one for the bandwidth degradation. Each session consisted of eight trials (one trial consisted on one 50-word list on each ear), four "asymmetrical" trials in which impaired target words are presented on one ear only and four "symmetrical" trials in which impaired target words are presented in both ears.

Subjects are divided into two groups: for group A, impaired target words in asymmetrical situation are presented on the right ear whereas for group B impaired target words in asymmetrical situation are presented on the left ear. Table II summarizes the different trial conditions.

Therefore, for each type of degradation, the experimental design was a nested design with three fixed factors (factor Quality with four levels, factor Situation with two levels, and factor Ear with two levels).

The order of conditions is different for each subject. Half of group A and half of group B began with the MNRU session, whereas the other halves began with the Bandwidth session.

For each condition (type of degradation x degradation level x situation x ear), twenty reaction times RT (time between the end of the target word pronunciation and a keyboard touch pressing) were measured as well as eventual errors or absences of response.

Document downloaded by @DAEL, 2026-11-11 23:48:40 - personal use only

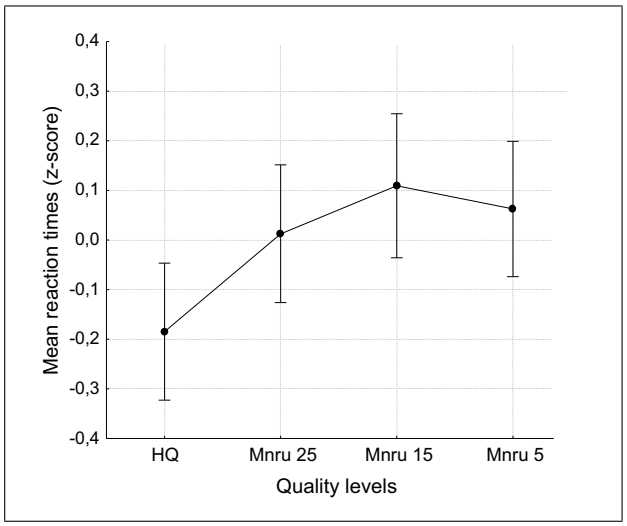


Figure 4. Averaged reaction times, as a function of quality levels.

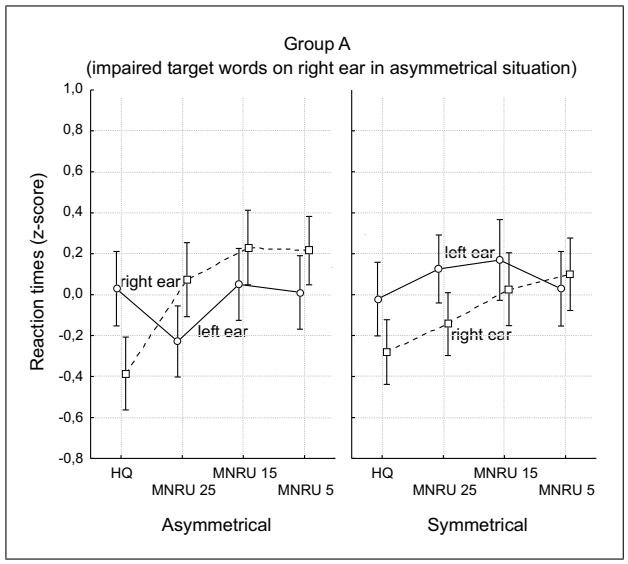


Figure 5. Z-score reaction times, as a function of quality levels, situations and ears, for group A.

During the whole test, the Galvanic Skin Response (GSR) as well the Heart Rate (HR) and the Blood Volume Response (BVP) were measured for each listening, with ProComp and the Biograph[1] software.

3.2.2. Results

In order to remove the inter-subject variability, reaction times are transformed into z-scores. Figure 4 shows the mean RTs (z-scores) and the associated 95-% confidence intervals as a function of MNRU levels. It seems that the worse the quality, the longer the reaction times.

An ANOVA on individual RTs considering the factors Situation (Asymmetrical / Symmetrical), Quality level (HQ High Quality, MNRU 25, MNRU 15, MNRU 5) and Ear (right or left), confirms the effect of quality ($F(3,714) = 18.56$, $p < 0.0001$). A Honestly Significant Difference Tukey test shows that the mean RT obtained for the HQ level is significantly shorter than the mean RTs obtained for the three other quality levels.

In addition, the ANOVA shows a significant effect ($F(1,238) = 37.054$, $p < 0.0001$) of the factor Ear. Figures 5 and 6 show the mean RTs and the associated 95-% confidence intervals according to the quality level, the situation and the ear, respectively for group A and group B. For non degraded conditions (HQ-HQ), it can be seen that mean RTs for the right ear are shorter than for the left ear, whatever the situation. The same observation can be done for degraded conditions, in the symmetrical situation (see Figures 5 and 6, right graphs). In asymmetrical situation, when degradations appeared on the right ear, the performances of this ear became similar to the left ear (Figure 5, left graph). In return, when degradations appear on the left ear, an increase of the difference between the two ears can be observed (Figure 6, left graph).

[1] Developed by Thought Technology (http://www.thoughttechnology.com/)

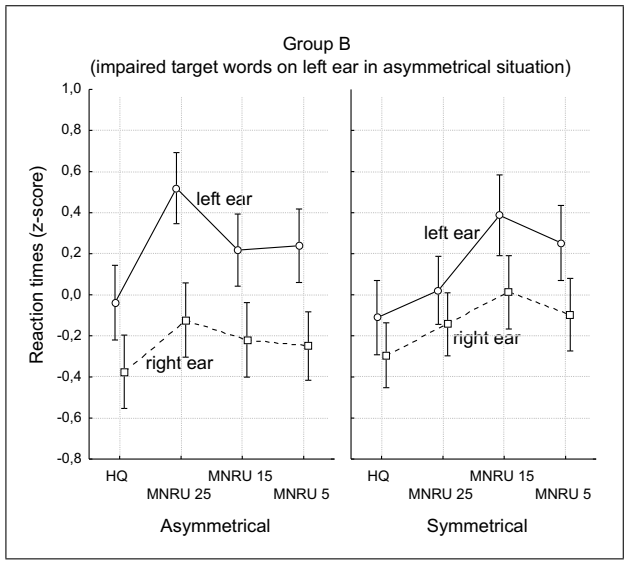


Figure 6. Z-score reaction times, as a function of quality levels, situations and ears, for group B.

Thus, right ear seems to be more effective, at least faster, than the left ear to detect the target words. Error rates confirm a predominance of the right ear on the left ear (less detection errors on the right ear) in the target words detection task in a dichotic listening situation. The predominance of the right ear for the word detection in dichotic listening might be the consequence by the Wernicke-Geschwind model that states that speech is predominantly processed in the left brain area connected to the right ear.

Similar effects are obtained for the bandwidths degradation (Quality effect: $F(3,714) = 3.03$, $p < 0.05$, Ear effect: $F(1,238) = 32.36$, $p < 0.0001$).As in the first experiment, no effect of quality level on the electrophysiological measures is found.

Document downloaded by @DAEL, 2026-11-11 23:48:40 - personal use only

3.3. Discussion

The result of these two studies show that speech quality, as it can be characterized by physical parameters (percentages of packet losses, MNRU, etc.) influences behavioural data such as reaction times or percentages of correct detection. However, this influence is not systematic and depends on the experimental protocols. For example, in the studied dual-task paradigm (cf. first experiment), speech quality only influences behavioural data for the secondary task. It highlights the extreme caution that has to be taken when building task-oriented protocols. These results are encouraging as they show that the influence of speech quality on behavioural data can be systematically assessed.

Contrary to Mullin *et al.* [14], our electro-physiological data were not influenced by quality. This absence of effect might be the consequence of a too low stress or motivation of our subjects resulting from the type of task they were assigned to achieve. In case of low stress or motivation, the degradation of signals in low quality conditions may not be have been considered by subjects as an obstacle hindering them from reaching an important goal. As a consequence, the electro-physiological data that should have reflected stress and negative emotion related to the difficulty of achieving the task did not significantly vary. The control of stress and motivation of subjects in task-oriented protocols will be discussed in section 6.

In our task-oriented protocols, speech quality is not the measured object, but a parameter that allows controlling the amount of difficulty to achieve the task. In the same way, one can hypothesise that in ecological situations, speech quality can be considered as an *enabler* of the communication, enabling the user to reach his goal if it is high, and hindering him/her if it is low. Another enabler is for example the ease of use of the terminal/service that might be assessed through usability tests. The role of this enabler (facilitation / hindrance) in a task- oriented protocol can by studied on the basis of the analysis of the subject's behaviour. Such approach leads us to consider speech quality in the framework of appraisal theory.

4. Appraisal theory in the context of voice telecommunications

The first theories on emotion, such as the *arousal theory* or the *discrete emotion theory* [21] mainly considered full-blown emotions (often referred as the big six after the work of Ekman [22]: anger, disgust, fear, sadness, surprise and happiness). These emotions are generally elicited by straightforward and strong stimulations. Typical examples are: seeing a snake triggers fear, eating aversive food triggers disgust, hearing a sudden loud sound behind triggers surprise, etc. Research on emotions in the past forty years has considerably refined the concept of emotion and its associated theories. Today, one of the most comprehensive and accepted emotion theory is the *appraisal theory*, which takes into account not only full-blown, but complex and blended emotions such as boredom, irritation, the mix of relief and fun, etc. [2, 23]. This theory considers that emotions are based on subjective evaluations of the signification of events which have a matter with the well-being of the concerned person. Lazarus and Folkman [24] have introduced the concepts of primary and secondary appraisals. The primary appraisals deal with the pleasantness/unpleasantness of the object/event under evaluation, and whether it helps or hinders satisfying a need or reaching one's goal. The secondary appraisals are related to what extent the person can cope with the consequences of the object/event under evaluation.

Full-blown emotions are considered as being associated to automatisms that can bypass cognition in certain conditions (e.g. seeing a snake => full blow emotion of fear => running away and shouting). Except these specific cases where little control seems to be left to the person who is subject to the full-blown emotion, it is generally admitted that emotions afford behavioural flexibility and play a central role in the regulation of cognitive and motor behaviour [25]. Emotions are also essential for triggering attention on and memorizing objects/events [26]. Actually, emotion components include most of all parts of psychological functioning (feeling, physiological changes, motor expression, action tendencies, cognitive processing). Emotion processing is therefore a complex mechanism that takes place in the whole organism (the body and the brain) and which leads to consider the specificity of an *emotion episode* (as opposed to a global psychological functioning) [2] only as a question of threshold in intensity and duration of cognitive, physiological and motor activity.

As discussed above, during a voice communication, users will generally not pay a specific attention to speech quality. However, for a given context (and the associated goals and expectations), if speech quality falls below a certain threshold, it will become a conscious object, grab their attention and trigger behavioural adaptations. From appraisal theorists' point of view, these behavioural adaptations can be considered as the consequence of the role that play emotions in the subjective evaluation of the signification of events affecting a person trying to reach a goal [2, 27]. In other words, if the enabler "speech quality" does not fulfill its role (enabling the user to communicate easily with his/her interlocutor), it will be appraised as an object that triggers an emotional reaction which in turn will trigger a (cognitive, physiologic, motor) reaction and a behavioural adaptation (concentrating, making the interlocutor repeat, moving, etc.). As a consequence, it might be argued that studying users' behaviours and their emotional reactions might be a valuable source of information on speech quality *and* on its impact on customer satisfaction.

5. Customer satisfaction

5.1. Customer satisfaction, resources consumption and expectations

During a voice telecommunication, to reach their goal, users will spend four types of resources:

- *cognitive resources:* perception, attention, information processing, reasoning, decision making, etc.

Document downloaded by @DAEL, 2026-11-11 23:48:40 - personal use only

- *physical resources:* physical resources required for cognitive processing, for speaking, for dialing and holding a handset, moving away from a noise source, etc.
- *telecommunication system's resources:* capacity of delivering and capturing a speech signal by the terminal, battery, display of information on the screen of the terminal, connection to the network, etc.
- *economic resources:* cost of the service for the user.

For a given context and a given goal, users have a certain expectation about the resources they need to spend to reach their goal and about the quality of the system they will use [3]. For example, if, at the top of a mountain, a climber wants to communicate with a mobile phone to his mother that he reached the summit, he might expect to encounter a poor availability of the service and a poor speech quality (because of the low network reception at this location) and a difficult context for communicating vocally (e.g. presence of noisy wind) that will affect his consumption of resources: first, he might try to phone his mother three times before getting the communication; second, he might need to concentrate hard on her voice because of the low quality and to shout because of the noisy wind around. Moreover, because of the low network reception, the consumption of the battery power of his mobile phone might also be important. Despite this high consumption of resources and the low speech quality encountered, if the climber reaches his goal, he will certainly feel very happy and rewarding to the telecommunication service. Let us take the same person, wanting to phone his mother for a little chat, sat in his sofa, at home, with a typical Public Switched Telephone Network (PSTN) service and a cordless handset. His goal, context and expectations will be completely different and he will certainly not tolerate to encounter such a low speech quality and to spend as many resources as in the mountain scenario. It is probable that at the end of his chat, he will not be specially rewarding to the telecommunication service. Moreover, in the mountain scenario he expected speech quality to be low, consciously felt it was the case and adapted his behaviour in consequence. On the opposite, it is highly probable that in the sofa scenario, he did not even consider speech quality during the communication and that speech quality did almost not affect his general behaviour.

Between these two extreme scenarios, there is a wide variety of contexts, uses and encountered speech qualities by users that will determine the impact speech quality has on their behaviours, on the several resources consumptions and on their overall satisfaction of the service. From what has been discussed previously, it appears that if one wants to study in the lab the object "speech quality" as closely as possible to what this object means in ecological situations, it implies to conduct task- oriented tests, where speech quality will be a controlled enabler to reach a given goal, and in which subjects' behaviours (measured resources consumption and performance), emotions (expressed, measured and self-reported) and satisfaction will be recorded and analyzed. The development of such a methodology will be discussed in section 6.

5.2. A definition of customer satisfaction

ISO's 9001 definition of customer satisfaction [28] is: "customer's perception as to whether the organization [service] has met his/her requirements". This general definition does not specify whether the service has allowed the customer to fully or partially reach his/her goals, and which and how many resources he/she had to spend during the use of the service. If we go back to Lazarus' concepts of primary and secondary appraisals, we see that those appraisals can be applied to the use of a service and that they allow an extension of ISO's definition.

During a voice telecommunication, Lazarus' primary appraisals will concern the pleasantness/unpleasantness of the different objects/events that will encounter the user, and whether these objects/events will help or hinder him/her to reach his/her goal. Secondary appraisals will concern to what extent the user can cope with the consequences of these objects/events (e.g. how to manage a poor speech quality during an important commercial call on VoIP?) In Lazarus' view, the outputs of these appraisals are complex emotions that will regulate user's behaviour. We suggest here that the output of these appraisals also contribute to user's (customer's, if we consider the economic link between the user and the phone operator) satisfaction, since they are the reflection of his/her evaluation of the different elements of the service. The outputs of the appraisals are

- primary appraisal: pleasantness and contribution of the service to reach user's goal. Pleasantness refers to the hedonic judgment of the elements of the service. For example, a voice telecommunication service with a high speech quality might offer the same contribution to a user to reach his/her goals but might be considered as more pleasant if the speech is wideband rather than narrow band. The contribution of the service to reach user's goal refers to the usability of the service and to the degree of completion of his/her goal resulting from this usability. For example, a complex terminal or a low speech quality service will complicate the communication and will hinder the user to reach his/her goal, therefore necessitating a behavioural adaptation of the user and an increase of resources consumption.
- secondary appraisal: extent to which the user can cope with the consequences of the elements of the service under consideration. As discussed in paragraph 4.1., the consequences of the variations of speech quality will be an increase of resources consumption. For a given context and a given goal, the user will tolerate only to a certain extent to spend more cognitive, physical, economic or system resources to compensate a low speech quality.

Considering the above elements, we propose the following definition of customer's satisfaction for voice telecommunication services.

For a given context of use and a given goal, customer's satisfaction is the result of his/her appraisal of the service according to

Document downloaded by @DAEL, 2026-11-11 23:48:40 - personal use only

- *the degree of completion of his/her goal reached thanks to the service;*
- *the usability of the service and the amount of resources spent to reach this degree of completion;*
- *the consequences for him/her of the amount of spent resources to reach this degree of completion;*
- *the pleasantness of the elements of the service*

as regard to his/her expectations and to prior experience.

Speech quality influences the first, second and fourth item of this proposed definition. This influence might be assessed through user tests which will be discussed in the following section.

6. Methodological implications

In the growing context of heterogeneity (of quality, uses, terminals and pricing) described previously, considering the strong demand from marketing and network units for producing useful and reliable data on speech quality, and the three drawbacks affecting MOS that were presented in the second section, there is an urgent need to develop test methodologies which furnish data that

- are not dependent on the range of speech qualities under examination during the test;
- are obtained without explicit judgment from subjects on speech stimuli;
- better reflect the role that speech quality plays in users' global opinion on the quality and satisfaction of telecommunication services where voice quality is an important matter.

6.1. Task-oriented protocols and motivation

As discussed in the third section, in a voice telecommunication context, speech quality appears to be an enabler that will influence user's emotional reactions, behaviour and satisfaction. In order to evaluate and characterize these three dimensions, one needs to define and measure dependent variables that should vary according to speech quality levels under test. To obtain such dependence, it is necessary to involve subjects in tests where they can develop a high motivation and where speech quality appears to be an important aspect for them (even if not explicitly described as such by the experimenter).

The problem of motivation in user tests is fundamental. In usability tests, when new services or new functionalities are tested, it is firstly tackled through a careful selection of subjects: only current or potential customers of the service under test are recruited. Second, task-oriented protocols are used where subjects have to achieve tasks which are similar or close to those they want to achieve when using the service (or an equivalent for potential customers) in real life. Our concern here is to develop test methodologies where subjects are not explicitly asked to give their opinion on speech quality (at least during the test). Consequently, we have to define a test where the goal and the motivation are not explicitly but implicitly connected to speech quality.

Research on the psychology of emotion shares the same concern of manipulating implicit variables in user tests where task-oriented protocols are used and where a high motivation of subjects is required in order to obtain realistic and spontaneous expressions of emotions [29, 30]. There are two families of protocols emerging from these studies. The first one is based on using an initial personal motivation of users, e.g. learning a new language. Aubergé *et al.* [31] used such a motivation in the e-Wiz experiment in order to obtain multimodal expressions of emotions. An application for learning a foreign language was proposed to motivated users who were manipulated in order to express emotions (by specific feedbacks of the application), without knowing that the aim of the experimenters was not to help them to learn this foreign language but to collect a multimodal emotional corpus. Such a protocol requires subjects to come regularly to the lab and is time consuming.

The other family of protocols is based on gaming. The natural tendency of many subjects to find a challenge and a motivation in a game where there is a competition (against the system or other participants) is used to obtain spontaneous emotional reactions [32]. The main advantages of game protocols are that they can be really challenging for users, that many test conditions can be tested in a relatively short period and that performance (the degree of achievement of the proposed tasks) can be rather easily measured. One of their major drawbacks is that performance is also dependent on prior competences of users in gaming. Simulating a competition in which there is a financial reward for the winner can be an additional factor of motivation that is used in labs.

6.2. Methods for user's emotional reaction assessment

Emotion is a psychological construct that can only be accessed through the analysis of its expression by the person under consideration or by its own introspection. Research on emotion in the last decades has considerably intensified and has furnished a large range of methods for recording, analyzing and interpreting emotional signs expressed by humans. The first category of methods refers to the recording and the analysis of electro-physiological signals such as skin conductance, heart rate and blood volume pressure. During emotion episodes, these signals reflect the activity of the sympathetic system and allow an interpretation of the emotional arousal. However, the valence of emotions can not be determined by the sole analysis of electrophysiological signals and other information are necessary [15].

A complementary source of information can be to achieve audiovisual recordings of subjects during the test and to analyze them *a posteriori*, video (facial mimics, gestures, body postures) and audio (voice prosody, vocal quality, volume, lexical content) signals being a valuable source of information for characterizing humans' emotional states [33].

Document downloaded by @DAEL, 2026-11-11 23:48:40 - personal use only

Audiovisual recordings of subjects' experience (the camera recording their point of view) during the test can also be made. These recordings are played back to subjects just after the test in order to immerse them back to what they just experienced. During this play back, they are asked to comment the recording and describe precisely the emotional experience they had. This method of self-confrontation has been used e.g. by Le Chenadec, Maffiolo, Chateau and Colletta, [34].

A posteriori verbal reports can also be used. This method is less time consuming than the self-confrontation method, but cannot furnish a precise timing of subjects' emotional experience. The report can be free or guided, using a questionnaire.

Finally, pictorial methods are widely used, since the use of words for reporting emotional experiences presents several drawbacks [35]. Methods like the Self-Assessment Manikin (SAM, [36]) or the Product Emotion Measurement Tool (PrEmo [37]) are fast and easy to use, but again, can be only used *a posteriori* and furnish therefore only a global information on subjects' emotional experience, without a precise timing.

As pointed out in [38] and [39], no single method is sufficient and it is suggested to combine several methods to obtain a comprehensive and reliable analysis of subjects' emotional experiences during a test.

6.3. Methods for usability assessment

Several definitions of usability exist. ISO's 9241-11 [40] Guidance on Usability defines product usability as the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use. Nielsen [41], one of the most famous authors on usability, considers that the usability of a system is based on five dimensions: learnability (e.g. intuitive navigation), efficiency of use (e.g. time required to accomplish a given task), memorability (capacity of the user to recall information about the system after its use), errors (that should be few and easy to recover for the user) and subjective satisfaction. It has to be noticed that the definition used here for satisfaction is restricted to the subjective appreciation by the subject of the use of the system under test and is different from the concept of *customer satisfaction* that was discussed in the previous section.

Dependent variables measured during a usability test using a task-oriented protocol will be the degree of completion of the task (overall performance), the time and number of actions required to reach this degree of completion (efficiency), the progression of these variables throughout the test (learnability), the *a posteriori* recall by the subject of different information concerning the system (memorability), the number of errors made by the system and the way the subject could cope with them (error management) and diverse *a posteriori* judgments on these dimensions, including satisfaction, made by the subject and obtained in a post-test interview or questionnaire [41].

6.4. Methods for user's satisfaction assessment

Satisfaction is a rather complex concept to characterize. Whereas *customer satisfaction* encompasses several criteria such as usability, *user satisfaction* is considered by several authors as being one of the criteria of usability. Our aim is to characterize subjects' satisfaction towards the enabler "speech quality" when they are involved in a challenging game and that their performances in this game are directly dependent on speech quality. Speech might be embedded in a service (e.g. by using a communication tool, like in [42]) or might just be used as a medium to convey instructions on the game, like in [17]. In order to obtain more realistic data on subjects' overall satisfaction, it seems preferable to simulate the use of a communication tool. Most of marketing studies suggest the use of interviews, questionnaires or likert scales for assessing customer satisfaction [43]. It seems preferable to use such scales in order to obtain numerical data that can be correlated to emotional and behavioural data in subsequent analyses.

7. Conclusion

This paper suggests that speech-quality test methodologies should move further in order to take into account the two revolutions of mobile telecommunications and VoIP, and to treat several drawbacks of Mean Opinion Scores obtained in typical tests recommended by ITU-T [4]. It is proposed to consider valuable inputs from the research fields on the psychology of emotions, cognitive science, ergonomics and marketing research in order to develop test methodologies that will furnish more comprehensive, realistic and reliable data on speech quality and its role in voice telecommunication services. Introducing new concepts in the field of speech quality has the obvious disadvantage of complicating actual test methodologies (mainly based on simple listening tests) which are still widely used. No magic recipes are given and the authors try to contribute to the debate [17], [44], as others like [14] and [16]. We hope that this paper will stimulate discussions and reflections on the development of new test methodologies helping researchers and engineers to produce data that will be more useful for all the actors of the quality chain in the industry of telecommunications.

References

[1] ETSI-STQ (Speech Processing, Transmission and Quality Aspect): Future approaches to speech transmission quality across multiple interconnected networks. EG 201 474 V1.1.1, 2000.

[2] K. R. Scherer: Appraisal theory. – In: Handbook of Cognition and Emotion. T. Dalgleish, M. Power (eds.). Wiley, Chichester, 1999, 637–663.

[3] U. Jekosch: Voice and speech quality perception. Assessment and evaluation. – In: Signals and Communication Techology. Sringer, 2005.

[4] ITU-T Recommandation P.800: Methods for subjective determination of transmission quality. http://www.itu.int, 1996.

Document downloaded by @DAEL, 2026-11-11 23:48:40 - personal use only

[5] ITU-T Recommandation P862: Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. http://www.itu.int, 2001.

[6] G. A. Gescheider: Psychophysics: Method, theory, and application. 2nd ed. Lawrence Erlbaum Associates, Inc., 1985.

[7] V. Barriac, J. Y. Le Saout, C. Lockwood: Discussion on unified objective methodologies for the comparison of voice quality of narrowband and wideband scenarios. Workshop on Wideband Speech Quality in Terminals and Networks: Assessment and Prediction, Mainz, Germany, June 2004.

[8] ITU-T Recommandation P.800.1: Mean opinion score (MOS) terminology. http://www.itu.int, 2006.

[9] R. Guski: Psychological methods for evaluating sound quality and assessing acoustic information. Acustica united with Acta Acustica **83** (1997) 765–774.

[10] A. Gabrielsson, H. Sjögren: Perceived sound quality of sound-reproducing systems. J. Acoust. Soc. Am. **65** (1979) 1019–1033.

[11] F. Guyot, C. Piron, M. Castellango, B. Fabre: Characterization of the sound quality of vacuum cleaners. ASA 129th meeting, Washington DC, USA, May 1995.

[12] S. Namba: On the psychological measurement of loudness, noisiness and annoyance. A review. Journal of the Acoustical Society of Japan **8** (1987) 211–222.

[13] M. Merleau-Ponty: Phénoménologie de la perception. Gallimard, 1945.

[14] J. Mullin, L. Smallwood, A. Watson, G. M. Wilson: New techniques for assessing audio and video quality in real-time interactive communication. – In: Proceedings of IHM-HCI. J. Vanderdonckt, A. Blandford, A. Derycke (eds.). Lille, France, September 2001, 221–222.

[15] R. W. Picard: Affective computing. MIT Press, Cambridge, Massachusetts, 1997.

[16] G. P. Sonntag, T. Portele, F. Haas: Comparing the comprehensibility of different synthetic voices in a dual-task experiment. SSW3 (1998) 5–10.

[17] L. Gros, N. Chateau, A. Macé: Assessing speech quality: a new approach. Forum Acusticum, 4th European Congress on Acoustics, Budapest, Hungary, August 2005.

[18] ITU-T Recommandation P810: Modulated noise reference unit (MNRU). http://www.itu.int, 1996.

[19] A. T. Welford: Choice reaction time: Basic concepts. – In: Reaction Times. A. T. Welford (ed.). Academic Press, New York, 1980, 73–128.

[20] H. Pashler: Dual-task interference in simple tasks: Data and theory. Psychological Bulletin **116** (1994) 220–244.

[21] K. R. Scherer, T. Johnstone, G. Klasmeyer: Vocal expression of emotion. – In: Handbook of affective sciences. R. J. Davidson, K. R. Scherer, H. H. Goldsmith (eds.). Oxford University Press, New York, 2003, 433–456.

[22] P. Ekman: Basic emotion. – In: Handbook of cognition and emotion. T. Dalgleish, M. J. Power (eds.). Wiley, Chichester, England, 1999.

[23] N. Frijda: The emotions. Cambridge University Press, 1986.

[24] R. S. Lazarus, S. Folkman: Stress, appraisal, and coping. Springer, New York, 1984.

[25] K. R. Scherer: Cognitive components of emotion. – In: Handbook of affective sciences. R. J. Davidson, K. R. Scherer, H. H. Goldsmith (eds.). Oxford University Press, New York, 2003, 563–571.

[26] J. W. Schooler, E. E. Eich: Memory for emotional events. – In: Handbook of Memory. E. Tulving, F. I. M. Craik (eds.). Oxford University Press, New York, 2000, 379–394.

[27] P. C. Ellsworth, K. R. Scherer: Appraisal processes in emotion. – In: Handbook of affective sciences. R. J. Davidson, K. R. Scherer, H. H. Goldsmith (eds.). Oxford University Press, New York, 2003, 563–571.

[28] BS EN ISO 9001:2000: Quality management systems. http://www.iso-standards-international.com/, 2000.

[29] E. Douglas-Cowie, N. Campbell, R. Cowie, P. Roach: Emotional speech: towards a new generation of databases. Speech Communication - Speech and Emotion **40** (2003) 1–2.

[30] D. Ververidis, C. Kotropoulos: Emotional speech recognition: Resources, features, and methods. Speech Communication **1546** (2006) in press.

[31] V. Aubergé, N. Audibert, R. A.: Why and how to control the authentic emotional speech corpora. Eurospeech, Geneva, Switzerland, September 2003.

[32] K. R. Scherer: Emotion. – In: Introduction to Social Psychology: A European perspective (3rd. Ed.). M. Hewstone, W. Stroebe (eds.). Oxford, 2000, 151–191.

[33] D. Keltner, P. Ekman: Introduction: Expression of emotion. – In: Handbook of affective sciences. R. J. Davidson, K. R. Scherer, H. H. Goldsmith (eds.). Oxford University Press, New York, 2003, 411–414.

[34] G. Le Chenadec, V. Maffiolo, N. Chateau, J. M. Colletta: Creation of a corpus of multimodal spontaneous expressions of emotions in human-machine interaction. LREC 2006, Genoa, Italy, May 2006.

[35] A. Wierzbicka: Talking about emotions: Semantics, culture and cognition. Cognition and Emotion **6** (1992) 285–319.

[36] M. M. Bradley, P. J. Lang: Measuring emotion: the self-assessment manikin and the semantic differential. Journal of Behavioural Therapy and Experimental Psychiatry **25** (1994) 49–59.

[37] P. M. A. Desmet, P. Hekkert, J. J. Jacobs: When a car makes you smile: Development and application of an instrument to measure product emotions. – In: Advances in Consumer Research **27**. S. J. Hoch, R. J. Meyer (eds.). 2000, 111–117.

[38] R. W. Picard, S. Bryant Daily: Evaluating affective interactions: Alternatives to asking what users feel. CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches, Portland Oregon, USA, April 2005.

[39] N. Chateau, M. Mersiol: AMUSE: A tool for evaluating affective interfaces. CHI Workshop on Evaluating Affective Interfaces: Innovative Approaches, Portland Oregon, USA, April 2005.

[40] ISO 9241-11:1998: Ergonomic requirements for office work with visual display terminals (VDTs). http://www.iso-standards-international.com/, 1998.

[41] J. Nielsen: Usability engineering. Morgan Kaufmann, San Francisco, 1994. http://www.useit.com/.

[42] N. Chateau, V. Maffiolo, N. Pican, M. Mersiol: The effect of embodied conversational agents' speech quality on users' attention and emotion. ACII, Beijing, China, 2005, 652–659.

[43] B. E. Hayes: Measuring customer satisfaction: Development and use of questionnaires. ASQC Quality Press, Milwaukee, 1993.

[44] L. Gros, N. Chateau, V. Durin: Beyond the MOS. MESAQIN, Prague, Czech Republic, 2006.

Document downloaded by @DAEL, 2026-11-11 23:48:40 - personal use only